图书情报工作
第62卷 第13期 2018年7月

基于行为-内容融合模型的用户画像研究*

余传明[1] 田鑫[1] 郭亚静[1] 安璐[2]
[1] 中南财经政法大学信息与安全工程学院 武汉 430073 [2]武汉大学信息管理学院 武汉 430072

摘要：［目的/意义］为识别并去除非理性投资者的网络评论，提升评论的专业程度与质量，促进理性投资，本文以识别股吧中的用户是否属于噪声投资者为研究任务，进行用户画像。［方法/过程］对股吧的用户发文内容进行深度用户表示学习（deep user representation learning），结合股吧用户的粉丝数量、影响力、关注量、自选股、吧龄、发帖量、评论量、访问量等行为特征，提出一种行为-内容融合模型（behaviour and content combined model，BCCM），并在标注数据集上进行实证与对比研究。［结果/结论］实验结果显示，该模型对噪声投资者识别的F1值为79.47%，优于决策树方法（69.90%）、SVM方法（75.61%）、KNN方法（73.21%）和ANN方法（74.83%）。在噪声投资者识别这一特定用户画像研究任务中，通过利用深度用户表示学习引入文本内容特征，能够显著提升用户画像的各种评价指标。

关键词：用户画像 情感分析 用户表示学习 特征融合
分类号：TP391
DOI：10.13266/j.issn.0252-3116.2018.13.008

1 引言

随着智能手机与移动互联网技术的迅速发展与革新，人们的行为呈现明显的网络化趋势，网上用户行为数据以指数形式增长。从普通网页上的用户点击和浏览到社交平台上的转发与圈粉，从点评网站上的用户评论和点赞到电商网站上的用户购买和退换，网上行为数据呈现多样性、实时性、动态性、非结构化以及海量性等特征。如何合理、科学、有效地利用这些海量的网上用户行为数据，成为一个紧迫的现实问题。在这种情况下，用户画像的研究开始引起大数据分析领域相关学者的重视。

所谓用户画像，是根据用户人口统计学信息（demographic data）、社交关系（social network relationships）和行为模式（behavioral patterns）等信息而总结、抽象和挖掘出来的标签化用户模型。用户画像的早期研究是从商业角度出发，通过用户的消费习惯、消费金额、年龄、性别等特征判断用户的消费层次，从而进行精准营销。随着大数据技术的迅速发展，用户画像研究已经扩展到各个领域。例如，通过音乐平台用户的听歌习惯、听歌类型等行为数据来判断该用户所喜欢的歌曲类型（用户偏好画像），从而为其推荐相关的歌曲[1-2]；通过用户的体重、体质、血压、血糖、慢性疾病等指标来判断用户各项机能的健康与否（用户健康画像），从而推荐合理的膳食[3]；利用网站用户对不同内容的点击率、浏览时长等因素判断用户价值（客户关系画像），从而制定提高用户留存率的策略[4]。

值得说明的是，目前的用户画像研究较多地集中在利用用户行为特征上，对用户内容特征的深入研究则并不多见。鉴于此，本文以识别股吧中的用户是否属于噪声投资者这一特定用户画像任务作为研究目标，通过对股吧用户的发文内容进行深度用户表示学习（deep user representation learning），结合股吧用户的粉丝数量、影响力、关注量、自选股、吧龄、发帖量、评论量、访问量以及发帖长度等行为特征，进行实证与对比研究，以期为大数据环境下的用户画像提供借鉴。

* 本文系国家自然科学基金面上项目“大数据环境下基于领域知识获取与对齐的观点检索研究”（项目编号：71373286）和教育部哲学社会科学研究重大课题攻关项目“提高反恐怖主义情报信息工作能力对策研究”（项目编号：17JZD034）研究成果之一。
作者简介：余传明（ORCID：0000-0001-7099-0853），副教授；田鑫（ORCID：0000-0001-8929-7151），硕士研究生；郭亚静（ORCID：0000-0003-1443-8399），硕士研究生；安璐（ORCID：0000-0002-5408-7135），教授，博士生导师，通讯作者，E-mail：anlu97@163.com。
收稿日期：2018-01-04 修回日期：2018-04-02 本文起止页码：54-63 本文责任编辑：易飞

2 文献回顾

2.1 用户画像研究

用户画像研究较早应用于营销领域。赵曙光[5]通过深度访谈的方式对用户的社交媒体使用动机和行为进行综合提炼，将高转化率的社交用户提炼为5种类型，为针对不同类型的高转化率社交媒体产品设计、完善产品的用户体验、提高社交媒体营销效果奠定了基础。S. Yu 和 A. Gupta 等[6]利用社交网站 LinkedIn 中的用户数据，通过比较图摘要(graph summarization)与二分图(bipartite graph learning)两种方法来识别 LinkedIn 用户中的潜在购买者。M. Trusov 等[7]利用用户在线浏览数据来补充用户缺失信息，并利用仿真模拟算法验证其所提出方法在广告投放中的效果。I. Ha 等[8]利用社交关系网络来刻画用户画像，将结果用于广告投放之中，其对比实验表明，加入用户画像之后，广告投放的效果明显优于传统广告投放方式。

产品推荐与链接预测是用户画像应用的另一个热点领域。A. M. Elkahky 等[9]尝试利用深度学习对用户进行画像，并将其用于跨领域的用户产品(Windows Apps、新闻和电影/电视)推荐之中，其实验结果表明，针对普通用户，产品推荐的准确度能够提升49%左右。V. Codina 等[10]利用上下文感知推荐技术进行用户建模，并将其应用于旅游规划系统，对用户进行个性化旅游路线规划及推荐，通过以用户为中心的研究，验证了模型的有效性。T. Bensal 等[11]通过协作过滤的用户共同评估模式，结合新颖的分层贝叶斯建模方法对文章和评论内容进行主题建模，以向用户推荐可能感兴趣的新闻或博客文章，实验结果表明，该方法能够较好地解决推荐中的冷启动问题。G. Piao 等[12]提出结合知识库使用概念对 Twitter 等社交网络平台上的用户进行兴趣建模，实验结果表明，该方法在 Twitter 的链接推荐中能够显著提高准确度、召回率等各项评价指标。G. Piao 等[13]使用概念频率-逆文档频率(concept frequency-inverse document frequency, CF-IDF)作为用户建模策略，并融合了用户兴趣的时间动态和语义，将该策略应用于链接推荐预测，结果显示，该融合策略优于单策略的预测效果。

在移动互联网领域，用户画像也得到广泛应用。章成志等[14]通过收集用户手势行为，例如单击、双击、滑动、拖动和放大/缩小等对移动平台下 Web 阅读系统的用户画像进行了研究。黄文彬等[15]在利用某电信运营提供商的3万位在线用户记录数据，采用频繁项集挖掘等方法，从用户网络日志中所涵盖的位移信息构建移动用户行为画像。Y. X. Dong 等[16]利用手机互联网络(包括手机呼叫行为和短消息发送行为)对用户的年龄、性别画像，其识别准确接近80%。此外，用户画像还在用户评价预测[17]、入侵检测[18]、多媒体信息检索[19]等领域得到了较为广泛的应用。

2.2 噪声投资者识别研究

从噪声投资者识别的定义来看，它是用户画像技术在金融领域的典型应用。较早提出“噪声投资者”概念的是 A. S. Kyle，他将噪音交易者明确定义为无法获得内部信息，非理性地把噪音当作信息进行交易的投资者[20]。之后，J. B. D. Long 等将“噪音”概念模型化并提出了具有较强代表性的噪声交易者模型[21]，并将市场上的投资者分为理性投资者和噪声交易者。C. M. C. Lee 等将投资者情绪看作投资者在估计未来投资回报时，除公司基本面之外影响投资者判断的部分[22]。杨楷提出市场上存在两种投资者类型，一种是不受情绪影响的理性套利者，一种是易受外部情绪影响的非理性的噪音交易者[23]。E. M. Silva 和 L. Takimoto[24]提出一种新的模型对熟练的和非熟练的噪声投资者进行统一建模。文献[25]对上述内容进行了综合并指出，噪声投资者就是在做交易决策时，总是通过经验判断、过度自信、过度乐观或悲观、损失规避等一系列心理因素或认知偏差来做出决策，甚至是出现羊群行为(从众心理)的个人投资者。

从噪声投资者识别的方法来看，主要将文本信息作为判断噪声投资者的重要依据。例如，M. Rechenthin 等[26]根据雅虎财经的留言板信息(如发帖时间、内容长度、IP 地址等)找出潜在的噪声投资者，并进一步运用支持向量机、朴素贝叶斯等多种机器学习算法探索噪声投资者对股票走势的影响，结果表明加入噪声投资者因素对股票走势的预测更为准确，证明噪声投资者有一定影响力；L. F. Ackert 等[27]利用股吧论坛数据将影响力前10%的用户筛选为有影响力的投资者，认为他们的发言会对其他投资者造成影响且可信赖；T. H. Nguyena 等[28]用社会媒体的情绪预测股价走势，从留言板的文字中自动提取主题和相关情绪信息，对股票预测任务中情绪分析的有效性进行评估，结果在准确率上得到大幅度提升；S. Feuerriegel 等[29]基于对金融市场规则的认识提出利用文字新闻的交易策略，并提出基于监督和强化学习的自动化决策方法，将新闻数据一并纳入投资体系；池丽旭等[30]基于扩展卡尔曼滤波(extended Kalman filter, EKF)方法，构造出过

滤市场噪声的投资者情绪指标,实证结果表明情绪波动是影响资产定价的重要主观因素;熊伟等[31]实证检验股票特质波动率与股票收益和投资者情绪的相关性。研究发现,股票收益率对股票特质波动率的弹性,随投资者情绪的增加和噪声投资者比例的上升而增大。

从噪声投资者识别的应用来看,首先,噪声投资者影响着市场均衡和市场走势[32-33]。例如,辛荣等[34]通过分析噪声交易者情绪、信息质量和市场进化均衡状态的内生关系,发现在不同的噪声交易者情绪和信息质量下,市场会进化到与之相应的均衡状态;王宜峰等[35]构建情绪水平和变化综合指标,发现情绪变化对市场收益、市场风险均有显著正向影响;彭叠峰等[36]在单资产的两期定价模型中,分类出关注和疏忽两类投资者,在市场出清的均衡状态下发现提高信息关注度可以有效降低资产的风险溢价,并提出了"关注者分类假说"。其次,噪声投资者对于市场风险具有提示作用。例如,刘毅等[37]提出理性投资者也可能是风险偏好的,分析噪声交易策略及理性投资策略在金融市场的长期演化机制,结果表明在金融市场中,这两种投资策略有3种存在形式:收敛于噪声交易策略、收敛于理性投资策略、两者长期共存;V. Ramiah等[38]以区分新古典主义金融和行为金融为出发点,识别影响市场走势的异常事件,论证噪声交易与市场基本面之间的关联并建立模型定量化噪声交易风险;J. K. Shin和C. Subramaniane等[39]基于固定汇率和通货膨胀目标两项原则研究货币政策体系与噪声投资者在外汇市场中的关系,以托宾税为实例分析论证了在噪声投资者存在的情况下使用固定汇率的必要性,以及噪声投资者对外汇市场的影响能力。

值得说明的是,上述研究多数以评论或发帖文本(即内容特征)为主来刻画投资者,较少有将内容与行为特征融合以进行用户画像的方法。鉴于此,本文尝试将内容与行为特征融合,通过对股吧的用户发文内容进行深度用户表示学习(deep user representation learning),结合股吧用户的粉丝数量、影响力、关注量、自选股、吧龄、发帖量、评论量、访问量等行为特征,提出一种行为-内容融合模型,以期为用户画像研究提供借鉴。

3 研究方法

3.1 深度用户表示学习(deep user representation learning)模型

在噪音投资者识别这一特定任务中,用户内容特征主要来源于用户发帖及评论中的文本信息。我们将单个用户的所有发帖及评论文本整合在一起形成一个段落文本,所有的段落文本形成相应的语料库,利用该语料库进行用户表示学习。

本文所提出的用户表示学习方法受到词向量学习方法的启发,即:利用词向量来预测句子中出现的下一个单词。我们将这种思路应用到用户表示学习中,建立如图1所示的用户表示学习框架。在图1中,每个用户被映射到矩阵U中列表示的唯一向量,每个单词被映射到矩阵W中列表示的唯一向量。用户向量和词向量被平均或串联以预测语境(context)中的下一个单词。例如,利用用户向量u和"期待""股市"和"每天"3组词向量的平均值来预测下文是否会出现"上涨"一词(见图1)。在本文实验中,使用平均作为组合向量的方法,利用上述建立的语料进行训练以下相关的向量和参数。

(1)用户向量与词向量的获取。在模型启动阶段,用户向量和词向量被随机初始化,通过定义深度学习中的损失函数(即量化预测值与实际值之间的差距)和采用一定的优化方法(例如随机梯度下降方法,stochastic gradient descending),最终获得用户向量和词向量作为上述预测任务的间接产物。

(2)模型参数的获取。假设语料库中包含N个用户,词汇表中包含M个单词,我们想要学习用户向量,使得每个用户映射到一个p维向量,每个词语映射到一个q维向量,则模型总共有$N \times p + M \times q$参数。当N和M的值较大时,参数的数量可能也较大,参数更新在训练期间通常具有稀疏性。

利用语料库进行训练后,我们得到用户的内容特征,即用户向量。值得说明的是,相对于传统的特征工程方法,用户向量具有明显的优势,即:用户向量从未标记数据(unlabeled data)中学习,因此可以适用于没有足够标注数据的任务。用户向量的第二个优点是在小的语境中考虑到单词顺序,这点与n-gram模型方式相同,n-gram模型保留了段落的大量信息,包括单词顺序。由于传统的n-gram模型往往需要创建一个非常高维的表示,而用户表示学习模型能够创建一个相对低维的表示,因此用户表示模型相比于传统的n-gram模型,具有更好的推广性能。例如,可以将这些特征直接用于常规机器学习技术,如逻辑回归,支持向量机或K-means。在本文研究中,我们选择K-means算法将用户进行聚类,并将聚类簇分别标记为"0""1""2"等。

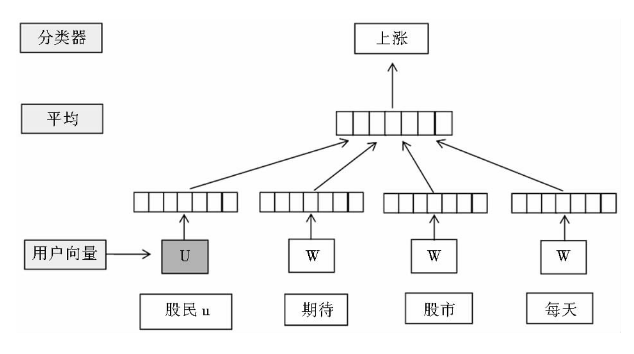


图 1　用户向量框架

3.2　行为与内容融合的 BCCM 模型

通过对股吧网站上用户信息的爬取，共得到 8 种行为特征：粉丝量、影响力、关注量、自选股、吧龄、发帖量、评论量、访问量。其中，粉丝量表示该用户被多少其他用户所关注；关注量是该用户关注其他用户的数量；自选股代表该用户所关注的股票的支数；吧龄代表用户在股吧上的活跃时期；发帖量是用户在股吧各个股票主题下所发布的总帖数；评论量是用户对所有帖子所发出的评论量总和；访问量代表该用户的个人主页被访问的次数；影响力是一个股吧提出的用于衡量用户影响力大小的指标，用 0－5 颗星来表示。选择上述行为特征主要基于以下考虑：一是特征具有可获取性，即能够通过爬虫软件以自动化方式获取；二是特征具有可用性，其效果在相关实证研究中得到了检验。例如，王凌霄等[40]采用关注数量作为用户参与程度的重要表征，将其应用到问答社区用户画像之中；林燕霞等[41]采用粉丝量作为微博群体划分的重要依据，并将其应用到基于社会认同理论的微博群体用户画像之中。

基于上述用户表示学习模型以及用户行为特征，我们提出一种行为－内容融合模型（behavior and content combined model, BCCM），用于识别股吧（http://www.guba.com）论坛上的噪音投资者。该模型基本步骤如下（见图 2）：

（1）首先针对用户评论及发帖的文本信息，运用 3.1 节深度学习方法获得用户表示，即用户向量（user embedding）；

（2）利用所获得的用户向量进行 K-means 聚类；

（3）将聚类标签作为一个特征加入到 8 个行为特征（即粉丝量、影响力、关注量、自选股、吧龄、发帖量、评论量、访问量）当中；

（4）将上述两类特征输入逻辑回归分类模型，最终识别噪音投资者。

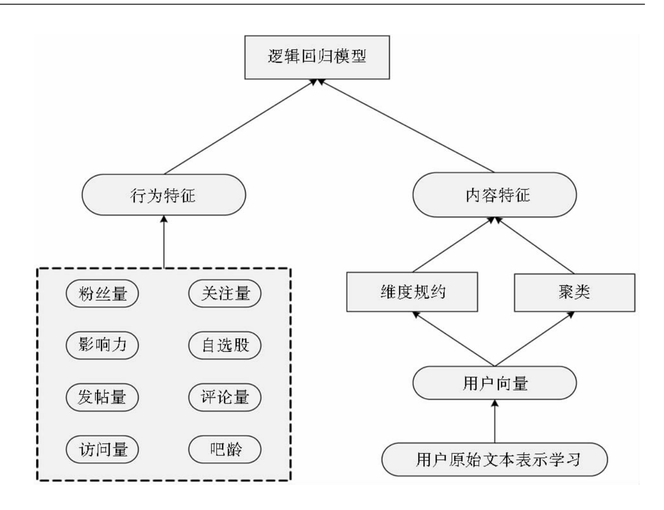


图 2　BCCM 模型的基本思路

4　实验

4.1　数据

本文的数据来自于东方财富网股吧论坛，参考其他相关社交媒体的用户画像领域相关研究[40－41]，结合股吧论坛所展示的信息，选择自选股数量、关注量、粉丝量、影响力、吧龄、总访问量、发帖量和回复量共 8 种数值型数据作为本研究的行为特征，用户的发帖及评论的文本信息作为内容特征。以“中金岭南”（网址为：http://guba.eastmoney.com/list,000060.html）为切入点爬取与 7 272 位用户相关的网络评论共计 10 万条，以及每个用户对应的自选股数量、关注量等行为数据。其中，用户发表评论所涉及到的时间跨度为 2008 年 8 月至 2017 年 3 月。

针对原始数据，我们从 7 272 位用户中随机抽取 2 500位用户，并对其进行人工标注。参与人工标注的为两名硕士研究生，二者系统地学习了金融信息工程、证券投资、金融学、宏观经济学、微观经济学等相关课程，持有证券从业资格证，在金融领域的噪声投资者识别方面具有一定的理论研究基础。

在标注过程中，当两人同时将某人标注为噪声投资者或理性投资者，则纳入我们的标注语料库中。如果一人将某个用户认定为理性投资者，而另一人有不同的标注结果，则将该用户数据从标注语料库中剔除，这类用户在 2 500 条数据中占比 15.4%（391 人）。对于标注不一致（即一人标注为理性投资者，而另一人标注为噪声投资者）的情况，尝试引入更多的标注者，但并不能有效地解决歧义问题。通过对原始语料进行观察，发现这些用户的评论和行为通常具备理性和噪声投资者的双重特点，标注者对此很难得到一致性的判断。鉴于此，为保证测试数据的有效性，将人工标注不

一致的数据从标注语料库中剔除。

本文所获取的标注语料库呈现高度非均衡性,在取得一致性标注的 2 109 条数据中,理性投资者为 158 个,占比 7.49%;噪声投资者为 1 951 个,占比 92.51%。在标注过程中,我们发现,有些用户在不同支股票下的论坛中发布几乎完全相同的帖子,或每个发帖内容都不是一句完整的话,帖子内容并无参考意义;还有一些用户的发帖内容多为抱怨、提问、表情,甚至辱骂,明显受到情绪因素影响。在上述情况下,标注者更倾向于将其标注为噪声投资者。相反,在另外一些情况下,发帖内容往往带有较少或几乎没有情感因素,其发帖及评论内容偏向于立足于实际的理性的客观分析,标注者更倾向于将其标注为理性投资者。

从方法角度来看,上述经验标准隶属于基于规则的研究方法,准确度高,但较多地依赖于人工判断;本文所提出的行为-内容融合模型隶属于统计机器学习方法,不依赖于人工制定的规则。鉴于此,本文未将人工标注的经验规则直接采纳入算法当中,而是通过经验规则构建测试数据集,在此基础上检验统计机器学习方法在噪声投资者识别中的效果。

4.2 基线方法

本文使用了 6 种基线方法,分别是:SVM[42]、朴素贝叶斯[43]、决策树[44]、KNN[45]、ANN[46] 以及逻辑回归[47]。我们分别利用 6 种基线方法对自选股数量、粉丝量、影响力、吧龄等 8 个特征构建分类器模型。

4.3 评价指标

本文选用 F1 值作为实验的主要评价指标。F1 值是统计学中用来衡量二分类模型精确度的一种指标,同时兼顾了分类模型的准确率和召回率。当准确率与召回率两个指标发生冲突时,很难在模型之间进行比较,此时则需要用到 F1 值。

除此之外,本文还用到了 F_β 分数,F_β 的物理意义就是将准确率和召回率这两个分值合并为一个分值,在合并的过程中,召回率的权重是准确率的 β 倍。F_1 分数认为召回率和准确率同等重要,F_2 分数认为召回率的重要程度是准确率的 2 倍。

在本实验中,从标注好的理性投资者(噪声投资者)中尽可能多地识别出理性投资者(噪音投资者)是本文实验的主要训练目标,从这个意义上讲,召回率的重要性高于准确率。为了突出召回率的重要性,本文增加 F2 值作为评价指标。

5 实验结果与讨论

5.1 频率分布统计

本文对收集到的 8 种行为特征(自选股数量、关注量、粉丝量、影响力、吧龄、总访问量、发帖量和回复量),利用标注数据集进行频率分布统计分析。

从评论量来看,如图 3 所示,10.24% 的用户完全没有过评论行为,43.80% 的用户评论量集中在 0 到 50 之间,有 19 位用户评论量超过 500 条。通过对标注数据进行观察,我们发现,理性投资者在平均评论数量上(81 条)高于噪音投资者(60 条)。通过对理性投资者和噪音投资者评论内容进行对比,发现理性投资者对于他人的发帖通常是较为理性的赞成或反对;对于噪音评论者而言,其评论内容通常毫无依据,逻辑性较差。

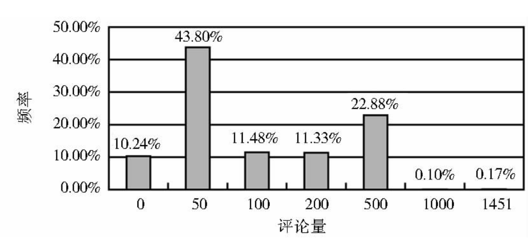


图 3 评论量频率统计

在发帖量方面,如图 4 所示,53.46% 用户的发帖量集中在 10 到 500 之间,仅有一位用户累计发帖量超过 40 万条。通过对标注数据进行观察,发现理性投资者在发帖平均数量上(855 条)高于噪音投资者(164 条)。通过对理性投资者和噪音投资者发帖内容进行对比,发现对于理性投资者,其发帖内容更具有金融领域的专业性以及参考价值;对于噪音评论者而言,其发帖内容(无论是自身发帖,还是对他人)通常专业性较弱,并且更倾向于较为夸张的辱骂和情绪宣泄。

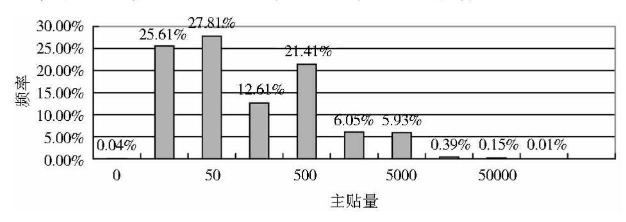


图 4 主贴量频率统计

关注量与粉丝量可以作为一组相对指标。图 5 是关注量频率统计直方图,可以看出,44.24% 的用户没有关注任何人,97.81% 的用户关注量不超过 50,仅有一位用户关注量超过 500。通过对标注数据进行观察,发现理性投资者在平均关注量(5 人)上低于噪

声投资者(7 人)。这说明理性投资者对外部不确定性信息的依赖性更小,没有过多关注噪声投资者的言论,理性投资者的决策行为主要受市场等客观因素影响。

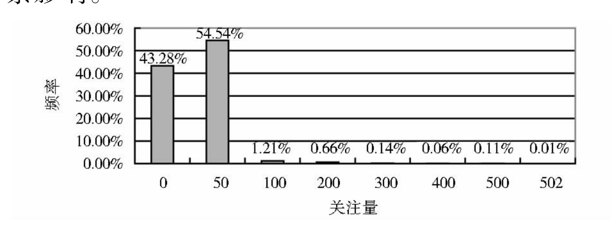


图 5　关注量频率统计

图 6 为粉丝量频率统计直方图,与关注量呈现出相似的分布结果:43.77% 的用户完全没有粉丝,85.75% 的用户粉丝量不超过 50 人,仅一位用户拥有百万粉丝(官方大 V)。通过对标注数据进行观察,发现理性投资者在平均粉丝量上(13 038 人)远高于噪声投资者(232 人)。这与常理相吻合,即:由于理性投资者更具有金融领域的专业性,其发帖内容更具参考价值,因此更能得到其他用户的认可,从而吸引到更多的粉丝。

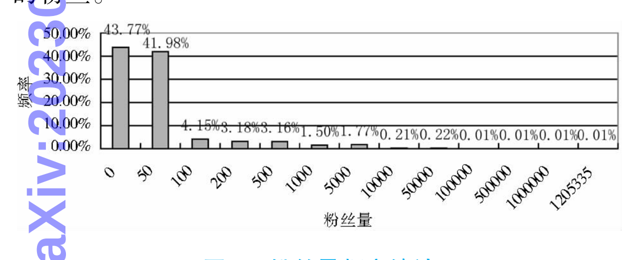


图 6　粉丝量频率统计

对比图 5 与图 6 的分析结果,发现理性投资者有着关注量低、粉丝量高的特征;噪声投资者与之相反。将平均粉丝量与平均关注量的比值作为阈值,在标注数据集中,发现,在用户的粉丝量与关注量的比值低于 33 时,用户是噪声投资者的概率显著高于理性投资者;在用户的粉丝量与关注量的比值高于 2 607 时,用户是理性投资者的概率显著高于噪声投资者。

5.2　基于行为特征的噪声投资者识别结果

将人工标注的数据集按 3∶1 的比例划分训练集与测试集,分别运用支持向量机模型(support vector machinese, SVM)、朴素贝叶斯(Naïve Bayes, NB)、决策树(decision tree, DT)、K 邻近算法(K nearest neighbours, KNN)、人工神经网络(artificial neural network, ANN)以及逻辑回归(logistic regression, LR)6 种基线方法针对 8 个行为特征构建二分类器,得到如表 1 所示的 F 值:

表 1　6 种基线方法的 F1 值结果展示

算法	P	R	F1	F_2
NB	3.75%	27.27%	6.59%	12.10%
DT	77.78%	42.42%	**54.90%**	46.66%
SVM	83.33%	30.30%	44.44%	34.72%
KNN	**100.00%**	9.09%	16.67%	11.11%
ANN	66.67%	36.36%	47.06%	40.00%
LR	46.00%	**51.00%**	49.00%	**49.91%**

由表 1 可以看出,在 6 种基线方法中,就准确度而言,KNN 算法取得了最好的效果(100%),排在其后的依次为 SVM(83.33%)和 DT(77.78%);就召回率而言,LR 取得了最好的效果(51%),排在其后的依次为 DT(42.42%)和 ANN(36.36%);就 F1 值来看,决策树取得了最优结果(54.90%),排在其后的依次为 LR(49.00%)和 SVM(47.06%)。考虑到召回率重要性高于准确率,对 F2 值进行了重点考察。表 2 可以看出,LR 取得了最高的 F2 值,达到 49.91%,高于 DT(领先 3.25%)、ANN(领先 9.91%)以及 SVM(领先 15.19%)。观察实验结果可以发现,在基于行为特征的噪声投资者识别中,LR、DT 和 KNN 这 3 种基线方法取得相对较优的综合效果。

鉴于 DT 在基线方法中取得了相对较优的综合结果,且相对其他算法而言,具有更强的解释性,本文尝试使用 DT 来进一步刻画噪声投资者的行为,其结果如表 2 所示:

表 2　噪声投资者行为刻画表

C						
C1:粉丝量 <20	Y	N	N	N	N	N
C2:评论量 >=72	-	Y	Y	Y	Y	N
C3:粉丝量 <700	-	Y	Y	Y	N	-
C4:自选股 >=13	-	Y	N	N	-	-
C5:主贴量 >=580	-	-	Y	N	-	-
A						
A1:理性投资者				√	√	√
A2:噪声投资者	√	√	√			

由表 2 可以看出,噪声投资者行为刻画因素主要分为 4 种,即粉丝量、评论量、自选股数量和主贴量(另外 4 种行为特征指标由于在实验表现中不够显著,未作为刻画指标)。对于噪声投资者而言,其行为画像倾向于以下 3 种情况:①粉丝量小于 20;②粉丝量位于[20,700)区间且评论量不小于 72,同时自选股数量不小于 13;③粉丝量位于[20,700)区间且评论量不小于 72,同时自选股数量小于 13,且主贴量不小于 580。对于理性投资者而言,其行为画像倾向于以下 3 种情况:

①粉丝量位于[20,700)区间且评论量不小于72,同时自选股数量小于13,且主贴量小于580;②粉丝量大于700且评论量不小于72;③粉丝量大于20且评论量小于72。

5.3 基于BCCM的噪声投资者识别结果

由于内容特征较能反映出用户的情感信息,字里行间隐含着其理性程度,使用深度用户表示学习方法获取了用户表示向量,利用K-means算法对用户的发帖及评论的文本内容进行聚类。将文本聚类结果加入到基线模型中,即选取9个特征(自选股数量、关注量、粉丝量、影响力、吧龄、总访问量、主贴量、评论量和聚类编号)对上述基线方法重新构建二分类器(即NB+C、DT+C、SVM+C、KNN+C、ANN+C),并与BCCM模型进行对比,得到如表3所示的实验结果:

表3 基线方法加入文本特征后的F值比较

算法	P	R	F1	F2
NB+C	1.35%	10.71%	2.39%	4.48%
DT+C	60.00%	32.14%	41.86%	35.43%
SVM+C	**83.33%**	17.86%	29.41%	21.19%
KNN+C	50.00%	7.14%	12.50%	8.62%
ANN+C	60.00%	42.86%	50.00%	45.45%
BCCM	71.00%	**74.00%**	**72.47%**	**73.38%**

由表3可以看出,在召回率方面,BCCM模型取得最高值(74.00%),远高于ANN+C方法(领先31.14%)、DT+C方法(领先41.86%)以及SVM+C方法(领先56.14%);在F1值方面,BCCM模型取得最高值(72.47%),远高于ANN+C方法(领先22.47%)、DT+C方法(领先30.61%)以及SVM+C方法(领先43.06%);在召回率方面,BCCM模型取得最高值(74.00%),远高于ANN+C方法(领先31.14%)、DT+C方法(领先41.86%)以及SVM+C方法(领先17.86%);在准确率方面,BCCM模型取值为71%,低于SVM+C方法,排在第二。综合4项指标来看,本文所提出的BCCM模型取得了最好的综合效果。这表明,通过深度表示学习加入用户内容特征,在样本非均衡的测试集合中,能够有效提升噪声投资者(以及理性投资者)识别的效果。

5.4 扩展实验结果

在5.3节中,所使用的标注数据集存在样本不均衡的情况(噪声投资者数量达到了92.51%之多,而理性投资者仅有7.49%)。为了进一步验证在均衡数据集中,BCCM模型相对于其他基线模型的效果优劣,我们采用过采样(欠采样)方法。具体步骤为,将2 109条数据按照3:1的比例划分训练集与测试集,在划分好的训练集中随机放回抽样1 500条数据作为新的训练集;同理,在测试集中随机放回抽样500条数据作为新的测试集,从而得到了一个较为均衡的数据集。在采样过程中,新的训练集和测试集中的数据可能重复出现。

利用重新采样过后的均衡标注数据集,我们重复了5.2节实验,得到实验结果(见表4)。表4中,字母O代表oversampling,即过采样。对比表1和表4可以发现,在采取过采样措施之后,各种基线方法的F1值得到了较为显著的提升,原本效果最差的朴素贝叶斯的F1值和F2值从10%左右提升到30%以上;决策树、支持向量机、K邻近及人工神经网络4种方法的F1值提升至80%左右;各种方法的F1最高值从原来的54.9%(见表1),提升到82.09%(见表4)。这表明,过采样方法对于提升各种基线方法的效果具有显著作用。换言之,表4中的各种基线方法在均衡数据集中具有更好的识别效果。

表4 运用过采样方法后的基线方法结果

算法	P	R	F1	F2
NB+O	26.63%	42.61%	32.78%	38.04%
DT+O	**89.56%**	70.87%	79.13%	73.96%
SVM+O	87.94%	76.09%	81.59%	78.20%
KNN+O	85.29%	75.65%	80.18%	77.40%
ANN+O	85.78%	**78.70%**	**82.09%**	**80.02%**
LR+O	74.00%	77.00%	76.00%	76.38%

利用重新采样过后的均衡标注数据集,我们重复了5.3节实验,得到实验结果(见表5)。表5中,字母O代表oversampling,即过采样;字母C代表content,即内容特征。由表5可以看出,在采取过采样措施以后,在召回率方面,BCCM+O模型取得最高值(81.00%),远高于KNN+C+O方法(领先9.70%)、ANN+C+O方法(领先10.57%)以及SVM+C+O方法(领先13.61%);在F1值方面,BCCM模型取得最高值(79.47%),高于SVM+C+O方法(领先3.86%)、ANN+C+O方法(领先4.64%)以及KNN+C+O方法(领先6.26%);在召回率方面,BCCM模型取得最高值(80.38%),高于ANN+C+O方法(领先8.25%)、KNN+C+O方法(领先8.32%)以及SVM+C+O方法(领先9.93%);在准确率方面,BCCM模型取值为78%,高于NB+C+O以及KNN+C+O方法。对比表3和表5可以看出,在采取过采样措施以后,尽管与其他基线模型的效果领先程度有所缩窄,本文所

提出的 BCCM 模型仍然取得了最好的综合效果。这表明，通过深度表示学习加入用户内容特征，在样本均衡的测试集合中，能够有效提升噪声投资者（以及理性投资者）识别的效果。

表5　运用过采样方法后的结果对比

算法	P	R	F1	F2
NB＋C＋O	27.78%	32.61%	30.00%	31.51%
DT＋C＋O	84.57%	59.57%	69.90%	63.31%
SVM＋C＋O	**86.11%**	67.39%	75.61%	70.45%
KNN＋C＋O	75.23%	71.30%	73.21%	72.06%
ANN＋C＋O	79.80%	70.43%	74.83%	72.13%
BCCM＋O	78.00%	**81.00%**	**79.47%**	**80.38%**

5.5　讨论

从各种算法的总体实验结果对比来看，在5.2至5.4节实验中，BCCM模型显示了较高的稳定性，该模型能较好地实现噪声投资者（理性投资者）识别，其实验结果优于传统的基线方法。该模型的提出在一定程度上推进了噪声投资者的自动化有效识别，通过区分噪声投资者与理性投资者并剔除掉噪声投资者的噪声信息可以为决策者提供引导性建议，具有一定的可靠性与参考价值。值得说明的是，BCCM模型利用未标记数据（unlabeled data）进行机器学习，相对于传统的监督式机器学习方法，能够节省繁重的人工标注任务，因而更加适用于没有足够标注数据的任务。另外，在BCCM模型的用户表示学习模块中，用户向量较好地考虑到了小语境中的单词顺序，这点与n-gram模型方式相同（n-gram模型保留了段落的大量信息，包括单词顺序）。相对于传统的n-gram模型往往需要创建一个非常高维的表示，用户表示模型能够创建一个相对低维的表示（例如文本为100维），因而具有更好的推广性能。此外，本文通过识别来刻画噪声投资者行为，即将行为特征作为监督学习的输入以训练分类模型，分类的结果反映投资者的不同类型，最后根据分类模型所学习到的规则来刻画噪声投资者。

从实验结果与实际情况的比照来看，BCCM模型在不同的评价指标下都得到较好的效果，在多数情况下能够较好地识别噪声投资者。实验仍存在少数与实际情况不一致的结果。例如，ID为“暗******吖”的用户被人工标注为理性投资者，而实验结果将其判定为噪声投资者。通过对原始语料进行比对，该用户的发帖或评论包含理性分析的成分，但由于在成文上缺乏专业性，且语言表达规范性较弱，因此被模型误判为噪声投资者。再如，ID为“mk******777”的用户被人工标注为噪声投资者，而实验结果将其判定为理性投资者。通过对原始语料进行比对，该用户所发帖子中存在较多的广告，但其内容相对专业规范，且粉丝量等行为特征与理性投资者相符，因此模型将其误判为理性投资者。

从算法的推广性来看，本文中的BCCM模型基于东方财富网对噪声投资者及理性投资者的识别而提出。模型并非局限在噪声投资者识别这一领域，本文所提出的行为与内容结合的思路，可推广到其他相关社交媒体的用户画像领域。例如，采用行为与内容结合模型，将用户的产品评价文本用于深度用户表示学习模型，结合发帖量、转发率、注册时间间隔等行为指标，能够刻画出该用户属于真实评论者或是虚假评论者；将用户购买商品后的评价文本用于深度用户表示学习模型，结合用户的月均购买量、好评率、差评率等指标及其与人均水平的比较，能够刻画出该用户属于理性消费者还是非理性消费者；将微博用户的博文用于深度用户表示学习模型，结合用户的粉丝量、关注量、日均发博数量等行为指标，对用户进行用户画像，能够刻画出其是否为僵尸粉。此外，本文研究成果对于大众情感是否能影响股价的相关研究具有一定的现实意义。例如，可以利用本文的实验结果，在提取基于评论内容所反映的情感特征时，去掉噪音投资者的情感，或者去掉理性投资者的情感，以检验哪些群体更有可能对股价产生影响。

本文实验存在以下局限性：①由于标注人员有限，本文人工标注的数据仅有2 100多条，数据量不够充分，后续将标注更多的数据以增加论文的说服性；②在人工标注的过程中，存在一些不确定性误差，有些用户确实难以判定是否为理性投资者，所以尽管是多人统一的结果，依然可能存在误判的情况。

6　结语

本文以金融领域的噪声投资者（理性投资者）识别这一特定的用户画像为研究任务，在深度用户表示学习以及传统的机器学习的基础上，提出了一种新的结合内容与行为特征的噪声投资者识别模型，即BCCM模型。为了验证该模型的有效性，本文在原始非均衡标注集和采样后的均衡标注集上进行了多组对比实验。对比实验结果表明，在非均衡数据集上，BCCM模型方法所取得的R、F1和F2值均远高于传统的决策树、朴素贝叶斯、逻辑回归、支持向量机等基线分类方法；在均衡数据集上，BCCM模型方法优于传统的基线分类方法。综合各项实验结果表明，在噪声投资者识别这一特定用户画像研究任务中，通过利用深度用户

表示学习引入文本内容特征，能够显著提升用户画像的各种评价指标。

在后续研究中，将加入社交网络中的节点与信息传输中的特征，例如股民之间的互相评论、互相关注等，来进一步优化模型，以得到更好的用户画像效果；此外，还会将噪声投资者和理性投资者的分类结果应用到股票预测中，以进一步验证噪声投资者对于股价波动的影响。

参考文献：

[1] FERWERDA B, SCHEDL M. Personality-based user modeling for music recommender systems[C]// Joint European conference on machine learning and knowledge discovery in databases. Berlin, Heidelberg:Springer, Cham, 2016:254 – 257.

[2] YAN M, SANG J, XU C, et al. A unified video recommendation by cross-network user modeling[J]. ACM transactions on multimedia computing communications & applications, 2016,12(4):1 – 24.

[3] 王智囊. 基于用户画像的医疗信息精准推荐的研究[D]. 成都:电子科技大学, 2016.

[4] 吴明礼, 杨双亮. 基于移动特征数据的内容推送技术研究与应用[J]. 计算机技术与发展, 2017, 27(9):155 – 160.

[5] 赵曙光. 高转化率的社交媒体用户画像:基于 500 用户的深访研究[J]. 现代传播 – 中国传媒大学学报, 2014, 36(6):115 – 120.

[6] YU S, GUPTA A. Identifying decision makers from professional social networks[C]// ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2016: 333 – 342.

[7] TRUSOV M, MA L, JAMAL Z. Crumbs of the cookie: user profiling in customer-base analysis and behavioral targeting[J]. Marketing science, 2016, 35(3): 405 – 426.

[8] HA I, OH K J, JO G S. Personalized advertisement system using social relationship based user modeling[J]. Multimedia tools & applications, 2015, 74(20):8801 – 8819.

[9] ELKAHKY A M, SONG Y, HE X. A multi-view deep learning approach for cross domain user modeling in recommendation systems [C]// International conference on world wide Web. Florence, Tuscany, Italy:International world wide Web conferences steering committee, 2015:278 – 288.

[10] CODINA V, MENA J, OLIVA L. Context-aware user modeling strategies for journey plan recommendation [M] // User modeling, adaptation and personalization. Berlin, Heidelberg: Springer international publishing, 2015:68 – 79.

[11] BANSAL T, DAS M, BHATTACHARYYA C. Content driven user profiling for comment-worthy recommendations of news and blog articles[C]// ACM conference on recommender systems. New York: ACM, 2015:195 – 202.

[12] PIAO G, BRESLIN J G. User modeling on Twitter with word net synsets and DBpedia concepts for personalized recommendations [C]// ACM international on conference on information and knowledge management. New York: ACM, 2016:2057 – 2060.

[13] PIAO G, BRESLIN J G. Exploring dynamics and semantics of user interests for user modeling on Twitter for link recommendations [C]// International conference on semantic systems. New York: ACM, 2016:81 – 88.

[14] 汪强兵, 章成志. 融合内容与用户手势行为的用户画像构建系统设计与实现[J]. 数据分析与知识发现, 2017, 1(2):80 – 86.

[15] 黄文彬, 徐山川, 吴家辉,等. 移动用户画像构建研究[J]. 现代情报, 2016, 36(10):54 – 61.

[16] DONG Y X, CHAWLA N V, TANG J,et al. User modeling on demographic attributes in big mobile social networks[J]. Acm transactions on information systems, 2017, 35(4):1 – 34.

[17] TANG D, QIN B, YANG Y, et al. User modeling with neural network for review rating prediction[C]// International conference on artificial intelligence. Palo Alto, CA, USA: AAAI Press, 2015: 1340 – 1346.

[18] PENG J, CHOO K K R, ASHMAN H. User profiling in intrusion detection: a review[J]. Journal of network & computer applications, 2016, 72(1):14 – 27.

[19] FARSEEV A, NIE L, AKBARIk M, et al. Harvesting multiple sources for user profile learning: a big data study[C]// ACM on international conference on multimedia retrieval. New York: ACM, 2015:235 – 242.

[20] KYLE A S. Market structure, information, futures markets, and price formation[M]//. International ag-ricultural trade advanced reading in price formation market structure & price instability. Boulder, Colorado, USA: Westview Press, 1984:45 – 64.

[21] LONG J B D, SHLEIFER A, SUMMERS L H, et al. Noise trader risk in financial markets[J]. Journal of political economy, 1990, 98(4):703 – 738.

[22] LEE C M C, SHLEIFER A, THALER R H. Investor sentiment and the closed - end fund puzzle[J]. Journal of finance, 1991, 46(1): 75 – 109.

[23] 杨楷. 投资者情绪与股市短期波动的关系研究——对 2015 年中国股市的考察[J]. 未来与发展, 2016, 40(10):62 – 67.

[24] SILVA E M, TAKIMOTO L. How to model noise traders investors using prospect theory[J]. Open access library journal, 2017, 4 (4):1 – 7.

[25] 孔东民. 中国股市投资者的策略研究:基于一个噪音交易模型[J]. 管理学报, 2008, 5(4):542 – 548.

[26] RECHENTHIN M, STREET W N, SRINIVASAN P. Stock chatter: using stock sentiment to predict price direction[J]. Algorithmic finance, 2014, 2(3):169 – 196.

[27] ACKERT L F, JIANG L, LEE H S, et al. Influential investors in online stock forums [J]. International review of financial analysis, 2016, 45(1):39 – 46.

[28] NGUYEN T H, SHIRAI K, VELCIN J. Sentiment analysis on social media for stock movement prediction[J]. Expert systems with

applications, 2015, 42(24):9603－9611.

[29] FEUERRIEGEL S, NEUMANN D. Evaluation of news-based trading strategies[C]// International workshop on enterprise applications and services in the finance industry. Berlin, Heidelberg: Springer, Cham, 2014:13－28.

[30] 池丽旭，张广胜，庄新田，等. 投资者情绪指标与股票市场——基于扩展卡尔曼滤波方法的研究[J]. 管理工程学报, 2012, 26(3):122－128.

[31] 熊伟，陈浪南. 股票特质波动率、股票收益与投资者情绪[J]. 管理科学, 2015(5):106－115.

[32] KHOLDY S, SOHRABIAN A. Noise traders and the rational investors: a comparison of the 1990s and the 2000s [J]. Journal of economic studies, 2015, 41(6):849－862.

[33] ZHANG X, ZHANG L. How does the internet affect the financial market? an equilibrium model of in-ternet-facilitated feedback trading[J]. MIS Quarterly, 2015, 39(1):17－38.

[34] 辛荣，张强，陈彬彬. 噪声交易者情绪、信息质量与市场多重进化均衡[J]. 系统工程, 2016(4):9－17.

[35] 王宜峰，王燕鸣. 投资者情绪在资产定价中的作用研究[J]. 管理评论, 2014, 26(6):42－55.

[36] 彭叠峰，饶育蕾，雷湘媛. 有限关注、噪声交易与均衡资产价格[J]. 管理科学学报, 2015, 18(9):86－94.

[37] 刘毅，李景华. 噪声交易者在金融市场的长期存在性研究[J]. 管理评论, 2012, 24(7):36－41.

[38] RAMIAH V, XU X, MOOSA I A. Neoclassical finance, behavioral finance and noise traders: a review and assessment of the literature [J]. International review of financial analysis, 2015, 41(1): 89－100.

[39] SHIN J K, SUBRAMANIAN C. Monetary policy and noise traders: a welfare analysis[J]. Journal of macroeconomics, 2016, 49(C): 33－45.

[40] 王凌霄，沈卓，李艳. 社会化问答社区用户画像构建[J]. 情报理论与实践, 2018, 41(1):129－134.

[41] 林燕霞,谢湘生. 基于社会认同理论的微博群体用户画像[J]. 情报理论与实践, 2018, 41(3):142－148.

[42] TONG S, KOLLER D. Support vector machine active learning with applications to text classification[J]. Journal of machine learning research, 2001, 2(1):45－66.

[43] MCCALLUM A, NIGAM K. A comparison of event models for Naive Bayes text classification[C]//AAAI-98 workshop on learning for text categorization. Palo Alto, CA, USA: AAAI Press, 1998, 62(2):41－48.

[44] LANDGREBE D. A survey of decision tree classifier methodology [J]. IEEE transactions on systems, man, and cybernetics, 2002, 21(3):660－674.

[45] ZHANG M L, ZHOU Z H. ML-KNN: a lazy learning approach to multi-label learning[J]. Pattern recognition, 2007, 40(7):2038－2048.

[46] ZHANG G, PATUWO B E, HU M Y. Forecasting with artificial neural networks: the state of the art[J]. International journal of forecasting, 1998, 14(1):35－62.

[47] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Additive logistic regression: a statistical view of boosting [J]. Annals of statistics, 2000, 28(2):337－374.

作者贡献说明:

余传明:论文构思、数据获取、深度用户表示学习模型实验、论文初稿撰写与修改;

田鑫:机器学习对比实验、论文初稿撰写;

郭亚静:机器学习对比实验、论文修改;

安璐:论文构思和修改。

User Profiling Based on the Behaviour and Content Combined Model

Yu Chuanming[1] Tian Xin[1] Guo Yajing[1] An Lu[2]

[1] School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073

[2] School of Information Management, Wuhan University, Wuhan 430072

Abstract: [**Purpose/significance**] To identify and remove online reviews from irrational investors, enhance the professional degree and quality of comments, and to promote rational investment, this article takes identifying whether the users on the Guba website belong to the noise investors as an example, and carries out a user profiling study. [**Method/process**] Deep user representation learning method was used to learn text information such as users' posts, then a behavior and content combined model was proposed with respect to behavior characteristics such as fans number, influence, bar age, post number and so on, and an empirical and comparative study was done on the annotated data set. [**Result/conclusion**] Experiment result showed that the BCCM model got the F1 score of 79.47%, which is superior to Decision Tree model(69.90%), SVM model(75.61%), KNN model(73.21%) and ANN model(74.83%). In the specific user profiling task of identifying noise traders, by using deep user representation learning method to obtain text content characteristics, the various evaluation metrics of use profiling can be remarkably improved.

Keywords: user modelling emotional analysis user representation learning characteristic fusion